Thus, there remains a need in the art for a mechanism to increase the diversity of polymeric biological molecules present in a library and to increase the proportion of members of that library having a desired activity.

SUMMARY OF THE INVENTION

Methods to create information rich libraries, that is libraries that contain a high fraction of biological polymers having a desired activity are disclosed. The information used to create these libraries can include: multiple sequence alignments, substitution matrices, three dimensional structure, and prior knowledge about the structure and/or function of the reference sequence from which the library is to be produced of from a homologous sequence in a related molecule.

Generally speaking, the steps towards the manufacture of the libraries of this invention include generating a probability matrix, generating a constraint vector, designing a substitution scheme based on the probability matrix and constraint vector. The substitution scheme has utility as produced, and can be used to construct a library based thereon. The library can then be screened and the members of the library characterized. Data mining techniques can be employed to characterizing the functional clones. Optionally, the characterization data can be used as information in a subsequent iteration of the method to obtain a molecule with even more desirable properties.

Additionally, combinations of the methods described herein can be made with other techniques such as family shuffling and/or systematic scanning approaches can be performed in any order and for any number of iterations to produce the products described herein; such combinations are within the scope of the invention. Also provided are vectors containing polynucleotides produced by the disclosed methods, host cells comprising such vectors, proteins encoded by such polynucleotides, and libraries of members so generated.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 is a graphical representation of the relationship between a probability matrix and a constraint vector of this invention. After a probability matrix is generated, a constraint vector can be applied to the matrix to determine which amino acid substitutions will be selected to test for their effect on a desired functionality. In this graphical representation, the residues for

which values calculated by the matrix rise above the constraint put on by the vector are candidates for the library.

Figure 2 is an alignment of the sequence of ampC proteins from seven different organisms.

DETAILED DESCRIPTION OF THE INVENTION

The prior art is replete with examples of techniques intended to improve the function of proteins and polynucleotides under defined conditions. One of the most well known examples utilizes crossover recombination or DNA shuffling. Diversity produced by DNA shuffling is limited to the parent sequences and random mutations.

The invention described herein can be used to introduce residues that are not contained in the parent reference sequence but that are still likely to preserve structure and function. Because a constraint of functionality is placed on the possible mutations, the fraction of inactivating mutations is minimized. This allows one to test higher mutation frequencies and increases the chance of finding useful double and triple mutations. For example, in a library of double mutants there is one chance per member to find interacting mutations. However, if one can generate a library of members of which 100% are active and contain 20 mutations per member then there are 190 possible pair-wise interactions between these mutations per member. In addition, the library will contain a large number of functional proteins with triple and higher mutations.

DNA shuffling recombines linear blocks of sequence. This places many amino acids into new environments at the same time because residues which are close in linear sequence are not necessarily close in three dimensional space. Conversely, computer shuffling techniques allow one to recombine residues which are close in three dimensional space. Thus, one can effect mutations in subdomains of the protein which are distant in linear sequence but close in structure, thus further increasing the chance to find interacting mutations.

Because DNA shuffling recombines linear blocks of sequence, beneficial mutations at one locus may be masked by detrimental mutations nearby. For illustration purposes only, Ballinger found that recruiting a furin residue into position 104 of *Bacillus amyloliquefaciens* subtilisin improved performance of the enzyme. However, recruiting a furin residue at position 107 abolished expression of the protein. Because these residues are very close, the chances of

having a crossover event between them using DNA shuffling is remote and the resultant protein would not be active (if present at all) even though it contained a useful mutation. Ballinger, *Biochemistry* **34**:13312 (1995); Ballinger, *Biochemistry* **35**:13579 (1996).

Benefits of the invention described herein include greater control of the complexity of the library. For example, if a large number of functional proteins are desired, the constraint matrix can be constructed to include fewer substitutions likely to lead to non-functional proteins. If more diversity is desired, the constraint matrix can be constructed to provide a lower constraint on the probability matrix.

Because a library that has a higher percentage of mutated and functional proteins can be constructed, fewer members of the library are needed to achieve a suitable number of possible useful proteins. In a particular embodiment, one may characterize the sequence and function of most or all members of a population, including non-functional proteins. Thus, in addition to obtaining useful proteins with a minimal number of screening assays, one is able to obtain information as to which mutations are detrimental to a protein. This information can then be used in a new constraint matrix, for example for another iteration.

Knowledge-based approaches can incorporate information from mutation of the reference sequence into the substitution scheme. Such information can be derived from intentional mutagenesis, either sporadic or systematic, or can incorporate information from naturally occurring mutations. Systematic approaches can include saturation scans where each residue of a protein is individually changed to each of the other 19 genetically coded amino acids and the resulting single mutants screened for the desired property, as well as deletion mutagenesis scans where one or more residues are deleted from the protein, insertion mutagenesis scans where one or more residues are inserted in the protein, and alanine scanning mutagenesis where each residue of the protein is systematically replaced with an alanine. Although systematic approaches provide the most information, any mutation which provides information about the protein's ability to tolerate a mutation affecting the desired property can be used.

Before the present invention is described in detail, it is to be understood that this invention is not limited to the particular methodology, devices, solutions or apparatuses described, as such methods, devices, solutions or apparatuses can, of course, vary. It is also to